Data Science do ZERO

Capítulo 06 - Machine Learning

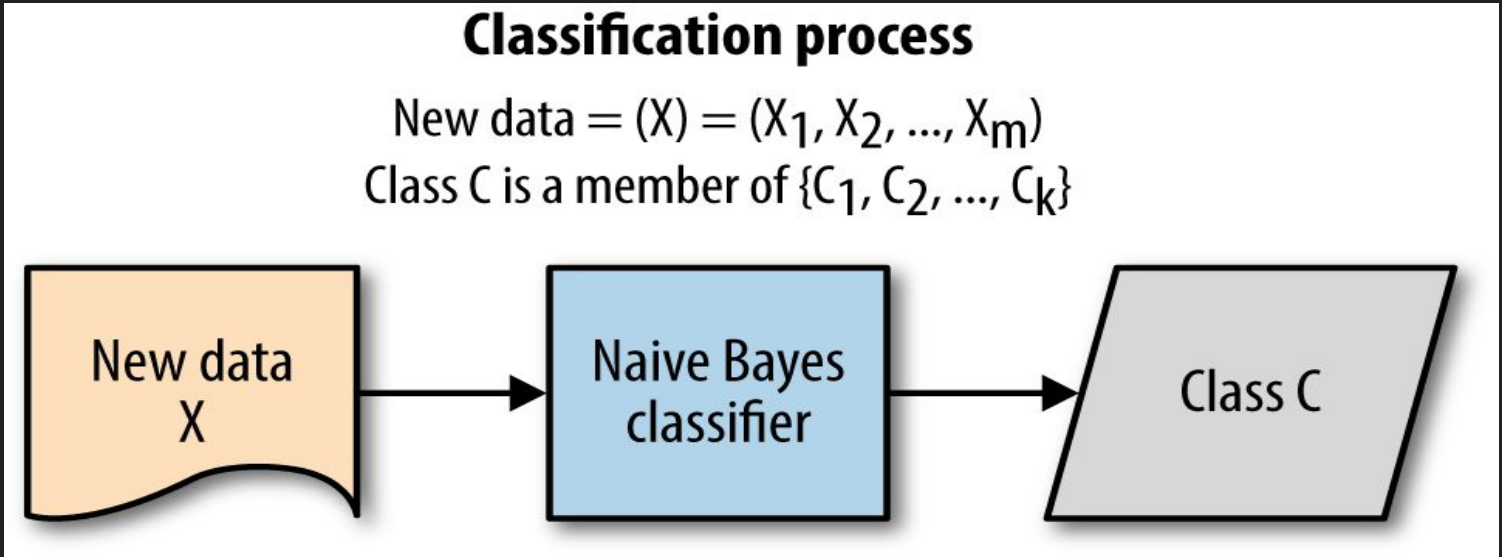
Naive Bayes

Naive Bayes

- Algoritmo utiliza aprendizado supervisionado.
- Classificador probabilístico baseado no teorema de Bayes.
- Assume que há uma **independência** entre as features.
- Por conta dessa caracteristica, recebe o nome de Naive.
- Possui poucos parâmetros e é um algoritmo simples e rápido.

Naive Bayes

- Fluxo da tarefa de classificação



Naive Bayes

- Baseado no teorema de Bayes.
- Probabilidade Condicional

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$$\text{posterior probability} = \frac{\text{conditional probability} \cdot \text{prior probability}}{\text{evidence}}$$

Naive Bayes

- Entenda o porque do nome “Naive”:

  I love my dog, but today it's terrible!!
- Essa frase poderá ter uma classificação positiva devido a independência de features.
- A probabilidade de cada palavra é independente das outras palavras.
- Essa independência de features é raramente encontrada em cenários reais.

Naive Bayes

- Cálculo probabilístico:
- Imagine um exemplo onde o objetivo é classificar algumas frases entre positivas ou negativas.
- Para isso eu tenho um dataset com frases e suas classes (positivo, negativo)
- Exemplo:
  - *“I love dog!”*, ‘positive’
  - *“Dog is bad in house”*, ‘negative’
  - *“The cat is love”*, ‘positive’
  - *“I hate cat and dogs”*, ‘negative’

Palavra	Classe
dog	positive
love	negative
bad	negative
love	positive
love	positive
dog	positive
house	positive
cat	positive
cat	negative
cat	negative
love	negative
house	positive
love	positive

Naive Bayes

- Tabela de frequência:
- O próximo passo é criar uma tabela de frequência por classe.
- Essa tabela é importante para o cálculo de probabilidades a seguir..

Tabela de Frequência		
Palavra	**Positive**	**Negative**
dog	2	
love	3	2
bad		1
cat	1	3
house	2	
Total	8	6

Naive Bayes

- Com a tabela de probabilidades das palavras e classe calcula-se as probabilidades de cada palavra para a base de dados.
- Quanto mais frequente a palavra for maior impacto no modelo.

Tabela de Probabilidades			
Palavra	**Positive**	**Negative**	
dog	2		2/14 = 0.14
love	3	2	5/14 = 0.35
bad		1	1/14 =0.071
cat	1	3	4/14 = 0.28
house	2		2/14 = 0.14
Total	8	6	
	8/14 = 0.57	6/14 = 0.42	

Naive Bayes

- Com a distribuição de cada palavra, é possível calcular a probabilidade de uma palavra pertencer a uma classe.
- Calculando a probabilidade da palavra “love” ser positiva ou negativa.
- A probabilidade da palavra “**love**” ser positiva é maior.
- **Importante:** Em casos de não haver a palavra na base, este retornará a probabilidade de da classe com maior frequência.

P(positive|’love’) = P(‘love’|positive) * P(positive) / P(‘love’)
P(negative|’love’) = P(‘love’|negative) * P(negative) / P(‘love’)

Calculando:

P(‘love’|positive) = 3/8 = 0.37 , P(positive) = 8/14 = 0.57, P(‘love’) = 5/14 = 0.35

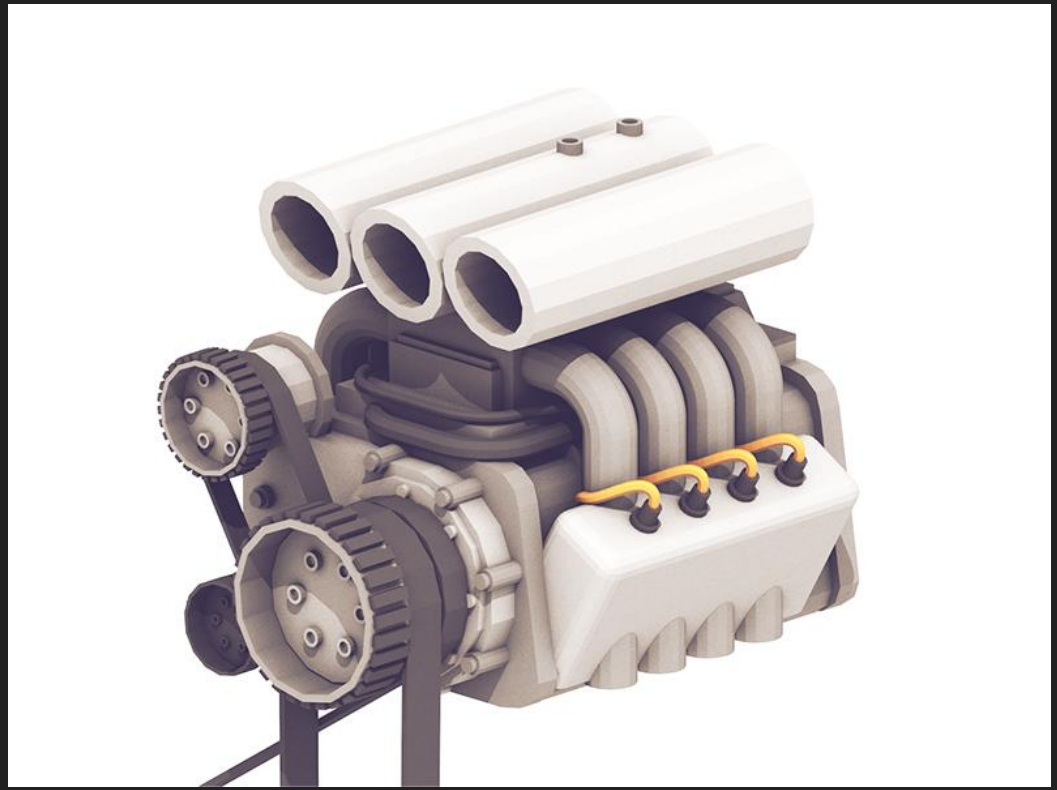
P(‘love’|negative) = 2/6 = 0.33 , P(negative) =6/14 = 0.42, P(‘love’) = 5/14 = 0.35

Agora, P(positive|’love’) = 0.37 * 0.57 / 0.35 = **0.60**

Agora, P(negative|’love’) = 0.33 * 0.42 / 0.35 = **0.39**

Naive Bayes

- **Multinomial Naive Bayes**:
  - Utiliza a frequência de termos
  - Muito utilizada em tarefas de classificação de textos.
- **Bernoulli Naive Bayes**
  - Variação do Algoritmo para valores para valores binários.
  - Trabalha com a matriz de presença de valores. Exemplo (1, 0)
- **Gaussian Naive Bayes**:
  - Variação do Algoritmo para valores contínuos.
  - Calcula-se a média e o desvio padrão dos valores de entrada para cada classe.
  - Assume-se que os valores estão em uma forma normal.



Hands on!